UNIVERSITEIT
GENT

FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Protecting Research and Technology from Espionage

Dirk Thorleuchter[1]

Dirk Van den Poel[2]

December 2012

2012/824

[1] Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany
[2] Ghent University, Faculty of Economics and Business Administration, Tweekerkenstraat 2, 9000 Gent, Belgium, http://www.crm.UGent.be

D/2012/7012/57

Protecting Research and Technology from Espionage

Dirk Thorleuchter[a,*], Dirk Van den Poel[b]

[a] Fraunhofer INT, D-53879 Euskirchen, Appelsgarten 2, Germany, dirk.thorleuchter@int.fraunhofer.de
[b] Ghent University, Faculty of Economics and Business Administration, B-9000 Gent, Tweekerkenstraat 2, Belgium, dirk.vandenpoel@ugent.be URL: http://www.crm.UGent.be

[*] *Corresponding author at: Fraunhofer INT, Appelsgarten 2, 53879 Euskirchen, Germany. Tel.: +49 2251 18305; fax: +49 2251 18 38 305*
E-mail address: Dirk.Thorleuchter@int.fraunhofer.de (D. Thorleuchter).

Abstract
In recent years, governmental and industrial espionage becomes an increased problem for governments and corporations. Especially information about current technology development and research activities are interesting targets for espionage. Thus, we introduce a new and automated methodology that investigates the information leakage risk of projects in research and technology (R&T) processed by an organization concerning governmental or industrial espionage. Latent semantic indexing is applied together with machine based learning and prediction modeling. This identifies semantic textual patterns representing technologies and their corresponding application fields that are of high relevance for the organization's strategy. These patterns are used to estimate organization's costs of an information leakage for each project. Further, a web mining approach is processed to identify worldwide knowledge distribution within the relevant technologies and corresponding application fields. This information is used to estimate the probability that an information leakage occur. A risk assessment methodology calculates the information leakage risk for each project. In a case study, the information leakage risk of defense based R&T projects is investigated. This is because defense based R&T is of particularly interest by espionage agents. Overall, it can be shown that the proposed methodology is successful in calculation the espionage information leakage risk of projects. This supports an organization by processing espionage risk management.

Key Words: Latent semantic indexing, SVD, Espionage, Risk assessment.

1 Introduction

Espionage is defined as the access to sensitive information without obtaining approval by the holder of the information (Crane, 2005). It is organized by foreign intelligence services (governmental espionage) or by corporations (industrial espionage) (Reisman, 2006) and it is executed by human experts (agents) in a specific target field that are able to distinguish between mundane information and information that is relevant for own organizational purposes (Kaperonis, 1984).

Many countries become very attractive to foreign intelligence services for governmental espionage and to corporations for industrial espionage because they are in an excellent geopolitical situation and because they host a large number of high-tech companies (Ho, 2008). In general, these countries are based on an open and pluralistic society, which make gathering of information easier for espionage agents. Two of these attractive countries are the Federal Republic of Germany and United Kingdom. Both, the UK intelligence services MI5 (Jones, 2008) and the German Federal Ministry of the Interior (2011), report that espionage target fields of foreign intelligence services and corporations are the applied science research and technology (R&T) and the military (that includes defense based R&T).

In general, R&T projects are sensitive concerning espionage if they deal with a specific combination of technologies and application fields (Warner, 1994). An example for a technology is diamond as a substitutive technology for silicon carbide (SiC). In contrast to SiC, diamond has several advances in developing a high-frequency power transistor for radar and communication applications. Projects dealing with this technology and application field are sensitive e.g. in Germany because Germany has - besides USA - worldwide leading knowledge in this technology. Further, these specific radar and communication applications are of military and also of commercial interest and thus, the application field is an important part of the national research strategy in

Germany. A further example is the nuclear weapon application field. It is trivial, that research projects e.g. in USA and United Kingdom (UK) examining technologies in this application field are sensitive because compared to non-nuclear-weapon states, USA and UK have a leading technological knowledge and they have strategic interests in this application field, too. Overall, R&T projects of an organization are sensitive concerning espionage if the combinations of technologies and application fields standing behind the projects are of organization’s strategic interest and if a worldwide leading - or at least competitive - technological knowledge is gained from the R&T projects (Lee, Chang, Liu, & Yang, 2007).

A general methodology for measuring risks is risk assessment. The risk assessment methodology defines a concrete situation and it identifies a possible threat (also named hazard). Then, the risk for the concrete situation concerning the hazard is estimated and prioritized by human experts. The risk depends on the estimated costs of a potential loss and it also depends on the estimated probability that this loss will occur (Marhavilas, Koulouriotis, & Gemeni, 2011).

A good measure for the sensitivity of an R&T project from an organization concerning espionage is the risk of an information leakage within this R&T project (Brunnermeier, 2005; Matsui, 1989). It can be seen that the costs of an information leakage are high if the project deals with a (strategic relevant) combination of technologies and application fields that is of particularly high interest concerning organization’s R&T strategy (Crawford & Sobel, 1982). Thus, the combinations have to be compared to the strategic interests of an organization for each project to estimate project’s costs of an information leakage.

Besides this strategic aspect, the knowledge aspect has to be considered. The probability that an information leakage will occur depends on current scientific knowledge of an organization. If a worldwide leading knowledge is gained from an R&T project then the probability that an espionage attack occur is large because espionage agents are not able to acquire the knowledge elsewhere. Otherwise, if the knowledge is already worldwide distributed then espionage agents are able to acquire this knowledge from other sources (Ho, 2007).

An R&T project is characterized by a project description that consists of textual information written by R&T planners and scientists that are working within the project. The description consists of information about the used technologies and the application fields. Textual patterns are extracted from the collection of all project descriptions that contain this information. This is done by use of latent semantic indexing (LSI) to consider aspects of meaning in the information because they are written by different human experts using different formulations. Based on machine based learning, on prediction modelling, and on a set of training examples, textual patterns are identified that are characteristic for projects dealing with strategically relevant combinations of technologies and application fields. These patterns are used to estimate the costs of an information leakage for new projects from a test set.

A new web mining approach is introduced that identifies the worldwide available knowledge of each strategically relevant combination (Almeida, 1996). For this, search queries are built based on terms from each corresponding textual pattern. LSI is applied on the retrieved web pages to select these pages where the corresponding textual patterns occur. Based on the uniform resource locators (urls) of the web pages, the geographical distribution of knowledge can be shown (Gorman, 2002). This enables the estimation of the probability that an information leakage occur.

In a case study, we investigate the information leakage risk of defense based R&T projects because defense-related technologies are a valuable target for espionage (Orozco, 2012). German Ministry of Defense funds a large number of projects with a wide technological scope and its R&T strategy shows several combinations of technologies and application fields with strategic relevance. Based on the descriptions of these R&T projects, the proposed methodology is applied and the information leakage risk is calculated for each R&T project. As a result, R&T projects are prioritized. This helps German Ministry of Defense by processing an espionage risk management to protect their technological knowledge.

Overall, this paper investigates a new application field: the protection of R&T projects from information leakage by espionage. It proposes a new semantic classification and web mining approach based on the standard risk assessment methodology for identifying the information leakage risk of R&T-projects. This helps researchers, research planners, and governmental agencies - that are responsible for R&T espionage protection - by the identification and prioritization of sensitive projects concerning espionage. It also ensures that risk reduction measures can be implemented in all high-risk situations despite limited resources (Yucel, Cebi, Hoege, & Ozok, 2012).

2 Background

2.1 Espionage in applied science R&T

Foreign intelligence services and corporations are interested in several target fields of espionage. Two of them with relevance to this approach are the strategic economic strength and the military capability intelligence (Jones, 2008).

The strategic economic strengths consist of aspects related to production or manufacture processes as well as to important infrastructures (Whitney & Gaisford, 1999). Very important factors for economic strength are research activities and technology development (Jaffe, 1986). In the last years, reports of German state offices for the protection of the constitution have shown that espionage activity in German R&T increases more and more (German Federal Ministry of the Interior, 2011). The espionage agents are normally recruited from the science and technology academia (Sivanesan, 2011). They stem from foreign intelligence services and from corporations of the Russian Federation, of the People's Republic of China, and of the Middle East, Asia and North Africa countries (Jones, 2008). Thus, protecting national or corporation-internal research activities and technology development against foreign countries or competitors is an important task (Brunnermeier, 2005).

The military capability intelligence consists of several aspects related to the strength of an enemy army where new weapon systems are specifically in focus. Besides collecting information about weapon system capabilities, the technologies standing behind the weapon systems are also an interesting target for espionage. Here, the corresponding espionage agents are usually trained military technologists (Jones, 2008). They focus on the large number of worldwide processed R&T projects that have the aim to contribute to future weapon systems by examining these new technologies.

Both target fields show that applied science R&T is an interesting target for espionage in general. Further, it specifically can be assumed that R&T in the area of security and defense are highly sensitive concerning espionage. Especially the rising asymmetrical threat during the last years forces governments to increase investments in security and defense related research and technology (R&T) (Gericke, Thorleuchter, Weck, Reiländer, & Loß, 2009). The European Union funds security research within the current European Framework Research Program (FP7) and it has founded the European Defense Agency (EDA) that funds defense based research with a wide technological scope among others (Oikonomou, 2012). European governments fund a large number of security and defense based R&T projects, too (Te Kulve & Smit, 2003). As an example, German Ministry of Defense coordinates over 1000 different simultaneously running technological research projects (Thorleuchter, 2008). Research projects in this area investigate technologies to apply them in a specific security and defence application field (Thorleuchter, Van den Poel, & Prinzie, 2010b). Thus, some of them may be a profitable target for espionage and they have to be protected from information leakage hazard (Thorleuchter & Van den Poel, 2011c; Thorleuchter, Weck, & Van den Poel, 2012a; Thorleuchter, Weck, & Van den Poel, 2012b).

2.2 Risk Assessment to measure espionage risk of R&T

The assessment of risks is the first step in risk management (Si, Ji, & Zeng, 2012). The aim of risk assessment is to measure the effect of uncertainty on critical assets. A recognized hazard is identified, characterized, and assessed. Then, the vulnerability of critical assets in a concrete situation is assessed concerning the hazard. Based on this assessment, the costs of a potential loss of the critical asset are estimated as well as the probability that the loss will occur.

Existing studies applying risk assessment have shown that it is difficult to measure the quantities for each risk (Cherp & Demidova, 2005). This is because both aspects, the costs of a potential loss and the probability that the loss will occur can only be measured intuitively by human expert estimations. The estimations can lead to a large chance of error. A further disadvantage is that estimations have been made for the effect of each hazard on each critical asset. That makes the estimation time-consuming for human experts especially in situations where a large number of hazards or a large number of critical assets (e.g. R&T projects) occur. As shown in Sect. 2.1, the rising asymmetrical threat causes the processing of a large number of R&T project in security and defense. Thus, the espionage risk of R&T projects especially in security and defense should be calculated by a quantitative estimation based on an automated approach.

Literatures propose qualitative risk assessment methodologies that investigate the information leakage risk of a small number of R&T-projects concerning governmental or industrial espionage by human experts (Thorleuchter, 2004; Thorleuchter & Van den Poel, 2012f). These qualitative methodologies also consist of the two criteria: costs of a loss and probability that the loss will occur as described below:

The costs of potential information leakage within R&T projects are estimated by considering strategic aspects. In an R&T strategy the critical milestones are described for realizing organization's R&T goals. The milestones are reference points (events) in an R&T project and the technologies that are necessary to achieve the critical milestones are prioritized by the R&T strategy. Thus, technologies that are examined for a specific application field are prioritized because they contribute stronger to organization's R&T goals than others (Solan & Yariv, 2004). Depending on organization's R&T goals (e.g. to be the first that introduces an innovative product in market), the costs of a potential information leakage within an R&T project examining a prioritized technology are higher than that of an R&T project examining a non-prioritized technology. Both qualitative methodologies suggest using the category '+' for R&T projects where technologies are applied in application fields that are prioritized by the strategy and using category '-' otherwise.

The probability that this information leakage will occur is estimated by considering the technological knowledge gained from the R&T project and by comparing it to the existing knowledge worldwide. If a technology is a unique feature offered by an R&T project then the project is very interesting for espionage (category 'A'). If the acquired knowledge also is available by other organizations and nations then the probability that an information leakage within the corresponding R&T project occur is medium (category 'B'). Otherwise, if the acquired knowledge is non-competitive (category 'C') that mean in many competitive organizations or nations a better founded knowledge is available then the corresponding R&T project normally is not an interesting target for espionage.

The qualitative methodologies calculate the risk by multiplying the costs with the probability. Based on the category combinations of the two criteria (furthermore they are named label) A+, A-, B+, B-, C+, C-, each R&T project is assigned to a label based on the qualitative estimation by human experts. The order of the labels concerning the risk of projects has to be determined based on organization's strategy e.g. the decision if the information leakage risk of a ‘A-‘ labeled project is greater than that of a ‘B+’ labeled project or not.

The proposed quantitative methodology in this paper uses the categories of the above mentioned qualitative risk assessment methodologies. In contrast to the qualitative methodologies, LSI as binary classification technique and prediction modeling is used to assign projects to category ‘+’ or to category ‘-‘. Further, web mining is used to assign projects to category ‘A’, ‘B’, or ‘C’. Thus, the proposed methodology realizes an automated assignment that enables the calculation of the espionage risk of a large number of projects.

2.3 Characteristics of R&T project descriptions

The proposed methodology uses text classification to extract textual patterns from project descriptions representing technologies and application fields. For this task, the relationships between different technologies and between technologies and application fields have to be considered. The relationships are described by a large number of literature studies (Choi et al., 2009; Fleck & Howells, 2001; Herstatt & Geschka, 2002; Jiménez, Garrido-Vega, Díez de los Ríos, & González, 2011; Radder, 2009; Rubenstein et al., 1977; Subramanian & Soh, 2010; Thorleuchter & Van den Poel, 2012e). Important findings from these studies with impact on the proposed methodology in this paper are described below:

Technologies are a collection of several different research topics within a specific technological field. R&T projects dealing with the same technology probably examining different research topics. Then, their descriptions consist of different terms although they belong to the same technology. Thus, it cannot be guaranteed that prevalent textual features from a technology occur in all projects, which belong to that technology. Additionally, different technological fields overlap that means the description of a project uses the same terms as a different project although both projects belong to a different technology. For a text classification approach it is important to consider these synonyms (words with different spellings but with the same meaning) and homonyms (words with different meanings but with the same spelling).

Technology descriptions are characterized by a high percentage of terms that occur together within a textual pattern. This means that the occurrence of different technological terms is not independent. Thus, the selected classification approach should consider the dependency of terms.

R&T projects examine one or several technologies to apply them in one or several application fields. Literature indicates different technological relationships (Geschka, 1983; Kim, Toh, Teoh, Eng, & Yau, 2012; Thorleuchter & Van den Poel, 2011a; Yu, Hurley, Kliebenstein, & Orazem, 2012). A substitutive technology is able to replace a technology by creating an application e.g. a specific energy supply application can be realized by using solar cell technology, electrical battery technology, or electrical fuel cell technology. Thus, projects dealing with a substitutive technology are also strategically relevant for an organization if the corresponding technology and its application field is relevant. A textual pattern that represents such a strategically relevant combination should contain terms from the technology and also terms from its substitutive technologies concerning the application. Further technological relationships are integrative, predecessor, and successor technologies that occur together or in a succession during the process of realizing an application e.g. fuel and lubricants technology for creating a new power plant application. Thus, to create a textual pattern as described above, the semantic relationship between terms describing a technology (including its substitutive, integrative, predecessor, and successor technology) and terms describing the application field has to be considered (Geschka, Lenk, & Vietor, 2002).

2.4 Text Classification

Text classification assigns pre-defined classes to text documents (e.g. R&T project descriptions) (Finzen, Kintz, & Kaufmann, 2012; Ko & Seo, 2009; Lin & Hong, 2011; Sudhamathy & Jothi Venkateswaran, 2012; Thorleuchter, Van den Poel, & Prinzie, 2010a). Classes can be textual patterns that represent 'technology – application field' combinations and text documents can be R&T project descriptions. In text classification, knowledge structure approaches can be distinguished from semantic approaches. Examples for knowledge structure approaches are instance-based learning algorithms (e.g. k nearest neighbor classification), decision tree models (e.g. C4.5), simple probabilistic algorithms (e.g. naïve Bayes), and support vector machine algorithms (Buckinx, Moons, Van den Poel, & Wets, 2004; Lee & Wang, 2012; Shi & Setchi, 2012). These approaches are not able to create semantic generalizations and thus, semantic relationship between terms (see Sect. 2.3). Further, some of them are not able to consider dependencies of terms (e.g. naïve Bayes).

Semantic relationship between terms can be identified by computational techniques that use statistical procedures on eigenvectors (Jiang, Berry, Donato, Ostrouchov, & Grady, 1999; Luo, Chen, & Xiong, 2011). These techniques consider words that are in a project description as well as words that might be in these descriptions (Thorleuchter & Van den Poel, 2012c; Thorleuchter & Van den Poel, 2012d;). LSI is a well-known representative of these techniques. It identifies hidden semantic textual patterns from a document collection (Park, Kim, Choi, & Kim, 2012). These patterns consist of terms that are not mentioned explicitly in a single document but that are related within the whole document collection descriptions (Christidis, Mentzas, & Apostolou, 2012; Tsai, 2012). LSI also considers synonym and homonym aspects as well as the dependency of terms (Thorleuchter, Van den Poel, & Prinzie, 2012). Thus, LSI fulfills the requirements from Sect. 2.3.

3 Methodology

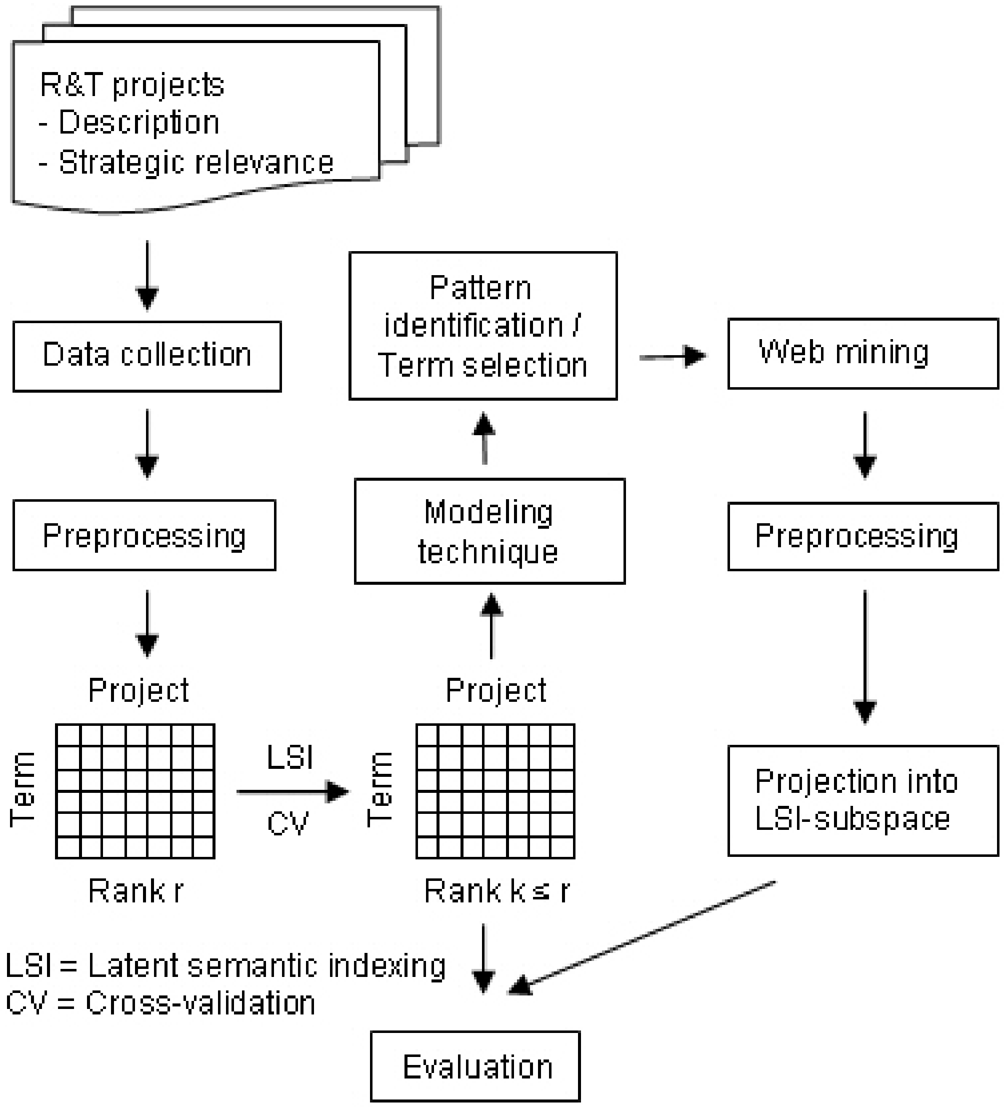

Figure 1: Processing of the approach

This new methodology investigates the information leakage risk of R&T-projects concerning governmental or industrial espionage. Sect 2.1 describes the espionage threat and the resulting information leakage risk for R&T projects in general and in particular for security and defence based R&T projects. The proposed methodology is based on a qualitative risk assessment methodology as described in Sect. 2.2. The relationships between technologies and application fields have an impact on the information leakage risk for R&T projects as described in Sect. 2.3. Thus, these relationships are considered in the proposed methodology by selecting a semantic text classification approach as described in Sect. 2.4 and by applying a web mining approach.

This methodology uses project descriptions of R&T-projects from an organization. Further, an estimation of human experts for each R&T project is used. It provides a binary assignment of a particularly strategic relevance based on organization’s strategic R&T interests. This means R&T projects with high contribution to the organization’s strategic R&T interests are distinguished from R&T projects with medium to low contribution to the organization’s strategic R&T interests. These assignments are used as target variable for prediction modeling and for the evaluation.

Projects are split in a test and training set. Their description is pre-processed by use of text mining methods. Based on the project description from the training set, a term-by-project matrix is constructed. LSI is applied to calculate hidden semantic textual patterns representing a combination of technologies and application fields. Prediction modeling with cross-validation is used on several rank k models to identify the value of k as the rank of the matrix with the optimal predictive performance. Semantic textual patterns are identified that frequently occur in the descriptions of R&T projects with particularly strategic relevance rather than in the descriptions of other R&T projects. They are used to predict the strategic relevance of an R&T project and thus, the costs of an information leakage. The test examples are projected into the same latent semantic subspace and each project is assigned to risk assessment category '+' if the identified patterns occur in the corresponding description. Otherwise, category '-' is assigned to the R&T project. This assignment is evaluated based on the estimations of human experts.

The identified patterns are selected and for each pattern several search queries are built containing a combination of several relevant terms. The search queries are executed and the received website addresses are sorted by different sections (e.g. country language code). After crawling the full text of the website addresses and after pre-processing, the results also are projected into the same latent semantic subspace as created during training. The number of results per section that contain the identified semantic textual patterns represents the knowledge distribution. As an example, many websites with a country language code .nl for the Netherlands that contain an identified semantic textual pattern show that the corresponding (particularly strategic relevant) knowledge is also available in the Netherlands. Based on the number of these results per section, each project is assigned to the category 'A' if a worldwide leading knowledge can be seen, to the category 'B' if the knowledge is available only in a few number of countries or corporations, or to category 'C' if the knowledge is distributed worldwide as described in Sect. 2.2.

3.1 Pre-processing

A pre-processing step is used to create a term vector in vector space model. It starts with text preparation where scripting code, tags, and images are removed. Typographical errors are corrected by using a dictionary. The text is split in terms (tokenization) and a conversion of terms in lower case is done. To reduce the number of different terms in the text, term filtering is applied. Stop words as well as terms that belong to a specific category (part-of-speech tagging) are discarded. Terms that appear only once or twice are also discarded based on Zipf's law (Zeng, Duan, Cao, & Wu, 2012; Zipf, 1949). Beside term filtering, term summarizing is also applied by converting terms to their stem (Thorleuchter & Van den Poel, 2012b).

Based on the pre-processing results, term vectors in vector space model are built for each project description. The size of the vectors depends on the different terms in the collection of project descriptions. Weighted frequencies are used for the vectors' components instead of raw frequencies because they improve forecast accuracy (Prinzie & Van den Poel, 2006; Prinzie & Van den Poel, 2007; Thorleuchter, Van den Poel, & Prinzie, 2010d; Van den Poel, De Schamphelaere, & Wets, 2004). To calculate the weight ($w_{i,j}$) of a term i in the R&T project description j, we use the formulation of Salton et al. (1994) that is based on term frequency ($tf_{i,j}$), on inverse document frequency ($\log(n/ df_i)$), and on a length normalization factor in denominator.

$$w_{i,j} = \frac{tf_{i,j} \cdot \log(n/df_i)}{\sqrt{\sum_{p=1}^{m} {tf_{i,j_p}}^2 \cdot (\log(n/df_{i_p}))^2}} \qquad (1)$$

3.2 Identification of semantic textual pattern with LSI

Each term vector created in the pre-processing step represents a project. A term-by-project matrix is constructed that contains all created term vectors. It consists of a large dimensionality that is unmanageable for further processing and has to be reduced. LSI (Deerwester et al., 1990) can be used together with singular value decomposition (SVD) to reduce the dimensionality by considering relationships between terms. This is done by grouping related terms to several semantic textual patterns. The patterns are selected based on their highest discriminatory power to other patterns. Thus, the reduced dimensionality can be determined by the number of these semantic textual patterns.

Mathematically, SVD splits the term-by-project description matrix A with rank r in three matrixes U, Σ, and V where U represents the impact of terms on the semantic textual patterns, V represents the impact of project descriptions on the semantic textual patterns, and Σ is a diagonal matrix containing the r positive singular values of matrix A ordered by size.

$$A = U \Sigma V^t \quad (2)$$

The weights for the matrix components of matrix A can be calculated by

$$w_{i,j} = \sum_{x=1}^{r} U_{i,x} \cdot \Sigma_x \cdot V_{j,x} \quad (3)$$

The dimensionality of matrix A is reduced by selecting a smaller value k that replaces the value of r. Based on formula 3, the weighted components for the reduced matrix A with rank k is calculated. The first k singular values are considered while the further singular values are discarded. Further, the first k columns of matrix U and the first k rows of matrix V are considered. As a result, the new k-rank approximation of matrix A with k semantic textual patterns is calculated by the product of the k-rank approximation of U, Σ and V:

$$A_k = U_k \Sigma_k V_k^t \quad (4)$$

The reduction of matrix A requires the selection of the value of k. k is the number of semantic textual patterns in the collection of all project descriptions. These patterns are used in a predictive model to classify projects as sensitive concerning espionage. The predictive performance of this assignment depends on the value of k. This is because a large value of k leads to the extraction of a large number of semantic textual patterns. Many of them cannot be used for prediction because they are probably irrelevant or unimportant for this task. Otherwise, a small number of k probably discards many relevant and important semantic textual patterns. This reduces prediction performance. We use an operational criterion as mentioned by Chen et al. (2010) for determining an optimal number of k. This criterion creates several rank-k models. It uses a parameter-selection procedure and a fivefold cross-validation to identify the rank-k model with the best predictive performance (Thorleuchter, Herberz, & Van den Poel, 2012; Thorleuchter & Van den Poel, 2012a).

The semantic textual patterns of the test examples have to be compared to those patterns created by the training examples and successfully used in prediction modeling. Thus, patterns from the test examples are projected into the same LSI-subspace as created by training (Zhong & Li, 2010). After pre-processing of test examples, each test example is represented by a term vector A_d as described in Sect. 3.1. A_d is used to calculate a new vector V_d for the project description d with

$$V_d = A'_d \cdot U_k \cdot \Sigma_k^{-1}. \quad (5)$$

Here, U_k and Σ_k are used as described in formula 4 and the components of the new vector V_d represent the impact of the project description on every semantic textual pattern as identified by training. Thus, this vector can be integrated into the matrix V_k.

3.3 Prediction Modeling

Prediction modeling is applied to predict an espionage sensitivity of R&T projects. The prediction is based on information about the espionage sensitivity of R&T projects in a training set. Logistic regression (Allison, 1999) is used to predict an espionage sensitivity of each project in a test set. The rationale standing behind logistic regression is easy to comprehend for decision makers (DeLong, DeLong, & Clarke-Pearson, 1988), the overall computing time is low, and it leads to robust results (Greiff, 1998).

The set $T = \{(x_i, y_i)\}$ consists of a concept vector for each project $x_i \in R^k$ and of a binary target variable $y_i \in \{0,1\}$. The concept vector represents the impact of a project on each semantic textual pattern as calculated in Sect. 3.2. A value of one for y_i indicates that the corresponding project is sensitive concerning espionage and a value of zero indicates that it is not. Based on a parameter vector w and on an intercept w_0 as calculated from a training set, logistic regression calculates the probability $P(y = 1 | x)$ that a project from a test set is sensitive for espionage by

$$P(y = 1 \mid x) = \frac{1}{1 + exp(-(w_0 + wx))} \qquad (6)$$

As a result, each project from the test set is assigned to a risk assessment category '+' if this probability exceeds a specific threshold.

3.4 Web Mining

The aim of the web mining step is to build and execute search queries to identify documents in the internet where the selected semantic textual patterns occur. For each of these semantic textual patterns, we identify five relevant terms that have an impact on this pattern above a specific threshold and that also have a term weight above a specific threshold (Thorleuchter & Van den Poel, 2011b). A search query is build for each pattern that consists of these five terms. The search queries are executed by use of Google as internet search engine. We use the web search advanced programming interface to enable an automatic execution of search queries and to enable an automatic processing of the results. Further, the Google translate advanced programming interface is used to enable an automatic translation of the terms in a search query from the target language (English) to several languages as available by the interface.

The search queries are executed and the received website addresses are sorted by different sections (e.g. country language code). The full text of the website addresses is crawled and the results are translated to the target language by use of the interface. After pre-processing (see Sect. 3.1), the textual documents also are projected into the same latent semantic subspace as created during training (see Sect. 3.2). For this, a concept vector is created for each document that represents the impact of the document on each semantic textual pattern (see Sect. 3.3). Further, prediction modeling is used to identify these documents, which contain information that is sensitive concerning espionage from the point of view of the own R&T strategy. These results are selected; the other are discarded for further processing.

The concept vectors of each R&T project from the test set are compared to all concept vectors from the selected results. A similarity measure is used to identify results with a similar impact on the semantic textual patterns and thus, with similar R&T activities.

The number of similar results (in total or split in different sections) is used for analyzing. A very small number of results show that worldwide the R&T activities in a project seldom occur and thus, the knowledge gained from the project is worldwide leading (category 'A'). A medium number indicates that further organizations also gained knowledge from these R&T activities. Then, the project can be assigned to category 'B' while a large number of results can be used to assign a project to category 'C'.

3.5 Evaluation criteria

This evaluation investigates the successfulness of the proposed automated approach that is based on a manual evaluation as done by human experts. The commonly used criteria are applied: the lift, the sensitivity, the specificity, and the area under the receiver operating characteristics curve (AUC).

The lift is a performance measure that is often used to measure classification performance for business applications. In this paper, it measures the increase in density in a specific percentile of the number of projects with high information leakage risk category relative to the density of all projects in total. For protecting projects from espionage it is important to increase the density of projects with high risk category especially in the top 10 to top 20 percentile because limited budgets and personnel resources for this task forces R&T planners to process risk reduction measures only for a small number of projects.

True positive (TP) is calculated by the number of projects with high espionage risk that are classified correctly while false negative (FN) is the number of these projects that are classified not correctly. The number of projects with low espionage risk that are classified correctly is true negative (TN) and false positive (FP) is the number of these projects classified not correctly. Then, the sensitivity is calculated by TP/(TP+FN) and the specificity is calculated by TN/(TN + FP). It is important to know that these criteria are varied if the value of the threshold is varied.

To be independent of a varied threshold, the AUC is used that is based on the receiver operating characteristic curve (ROC) (Migueis, Van den Poel, Camanho, & Cunha, 2012). It measures the area under a two dimensional plot as created by use of the sensitivity and the specificity criteria (Van Erkel & Pattynama, 1998). Hanley & McNeil (1982) show that for binary classification, the AUC is a successful performance measure.

4 Empirical verification

In a case study, we investigate the information leakage risk of defense based R&T projects funded by the German Ministry of Defense (GE MoD). Each project that has a long-term strategic relevance based on the R&T strategy of the GE MoD is classified as sensitive. Most R&T projects are processed to examine the potential of new technologies in general, to solve a current technological problem as occurred by the use of weapon systems, to improve the performance of existing weapon systems, or to satisfy current legislation (e.g. environment protection). These projects are not strategically relevant or they are at least of short-term strategic relevance. Thus, only 20% of all R&T projects investigate technologies in a specific application field that is of long-term strategic relevance for GE MoD.

Descriptions of more than 2000 R&T projects processed between 2000 and 2010 are selected. Defense based R&T is predominated by the United States of America (USA) because of their large defense budget. Thus from the GE MoD point of view, we define a worldwide leading knowledge in a technology (category 'A') as a knowledge that is worldwide unique by excluding knowledge available in the USA. Further, technological knowledge is defined as competitive (category 'B') if the technological knowledge of GE MoD is also available in other high-tech countries e.g. European Union member states, Russia, China, Japan etc. Otherwise, the corresponding R&T project is assigned to category ‘C’.

	Number of R&T projects	Relative percentage
Training set:		
Sensitive R&T projects	193	20
Non-sensitive R&T projects	772	80
Total	965	
Test set:		
Sensitive R&T projects	193	20
Non-sensitive R&T projects	772	80
Total	965	

Table 1: Characteristics of the data

In Table 1, the information about the training and test set is summarized. Both sets are randomly selected. The training set is used to calculate the semantic textual patterns and to estimate a regression model. The test set is used to evaluate the performance of the regression model. Evaluation results are compared to the frequent baseline as indicated by the relative percentage in Table 1.

4.1 Optimal dimension selection

To reduce the dimension of the term-by-project matrix, a cross-validated AUC is applied on the semantic textual patterns (dimensions) (see Fig. 2). It shows that the performance increases strongly up to the number of 50 patterns. From 50 patterns on, only a small increase in performance can be seen. A large number of patterns results in a large computational complexity. Thus, at 50 patterns an optimal point is reached concerning computational complexity and performance. The variable k is set to 50 and the examples from the test set are integrated into the latent semantic subspace that consists of these 50 semantic textual patterns.

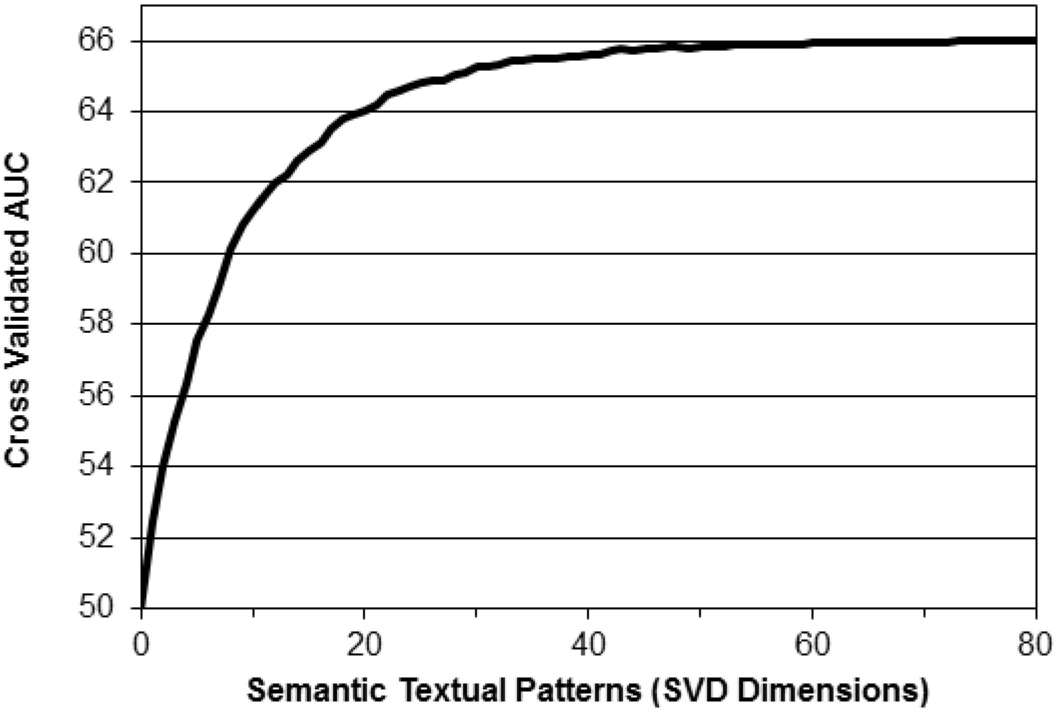

Figure 2: Calculating an optimal SVD dimension

4.2 Predicting espionage risk of R&T projects

The proposed methodology is applied and after the prediction modeling step, the R&T projects are assigned to category ‘+’ and ‘-‘. In the web mining step, the number of similar results is split in different sections (see Sect. 3.4). Sections are defined as the top-level-domain country code (ccTLD). The about 200 codes (e.g. .de for Germany) represent countries and the number of results per country is counted. The country codes for the USA are not considered. If the number of German results is much larger than the numbers of other nations then the corresponding R&T projects are assigned to category 'A'. If the number of German results is high and comparable with the numbers of some other countries then the corresponding R&T projects are assigned to category 'B'. Otherwise, if a low number of German results can be seen compared to the number of other countries then the corresponding R&T projects are assigned to category 'C'.

This research considers the fact that the number of websites in some top level domain country codes is larger than in other. As an example the German country code '.de' contains more websites than the Dutch country code '.nl' because much more people are living in Germany than in The Netherlands. Thus, a search query restricted on the German country code normally gets more results than the same (translated) search query restricted on the Dutch country code. For the calculation of the number of results, country weighting factors are used where e.g. a result from the '.nl' country code gets a larger weight than a result from the '.de' country code.

4.3 Comparing predictive performance

Based on the methodology of Thorleuchter (2004) (see Sect. 2.2), a manual assignment of R&T projects to the labels ‘+’ and ‘-‘ as well as to the labels ‘A’, ‘B’, or ‘C’ have been done by human experts from 2004 on. Thus, each of the 2000 R&T projects is manually labeled. However, risk reduction measures are processed only to the top 10 to top 20 percentile of projects with high risk category. Practically, all projects labeled with ‘A+’ are selected for further processing and none project that is labeled with ‘A-‘, ‘B+’, ‘B-’, ‘C+’, and ‘C-’. Thus, the assignment of further labels to the projects is not of interest. To consider this fact, only projects labeled with

'A+' are defined as projects with a high espionage risk and this definition is used as ground truth for the evaluation.

The cumulative lift curve in Fig. 3 shows the increase in density of all 'A+' labeled projects relative to the density of all projects in a specific percentile. In each percentile, the curve lies above the baseline that means the density of the R&T projects with high espionage risk is greater than the density of the baseline. However, it is more interesting to notice the large increase in density in the top 10 and in the top 20 percentile because this enables the selection of a small number of projects with high espionage risk for processing risk reduction measures.

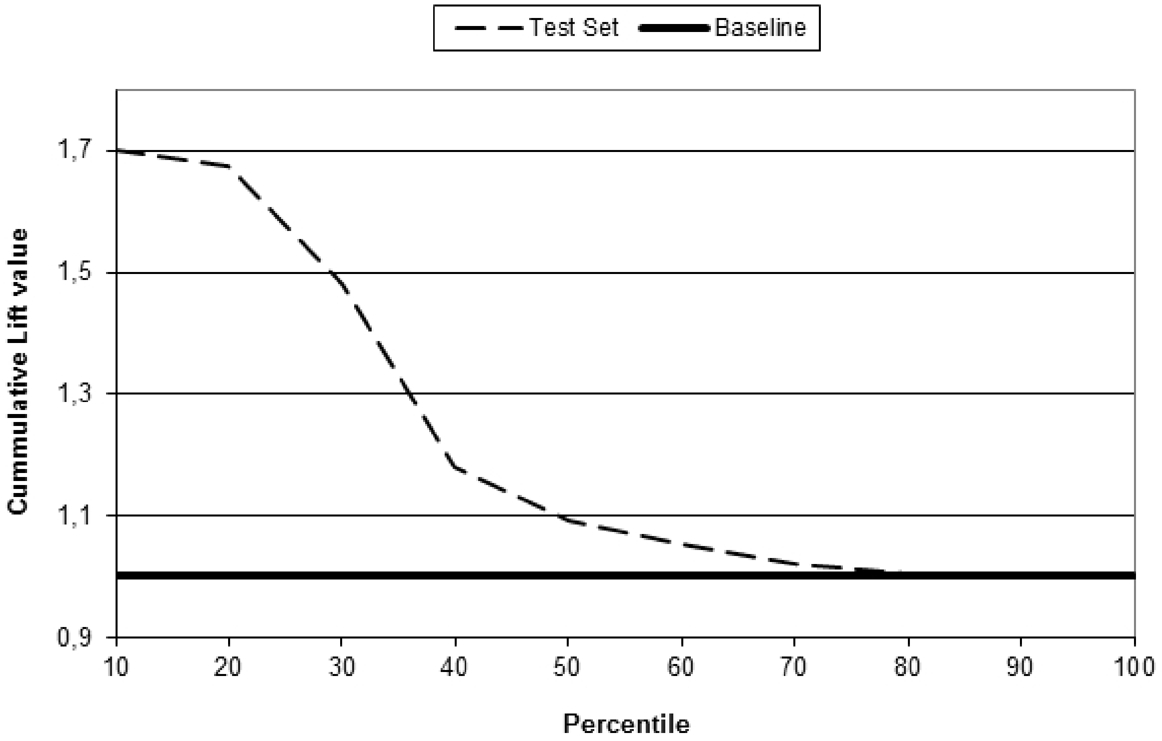

Figure 3: Cumulative lift value of the test set and of the baseline

The AUC is a performance measure for the binary classification of the proposed methodology (an R&T project is labeled with 'A+' or not) based on the ground truth as defined above. The ROC curve that is created by the processing of this case study lies above the ROC curve of the baseline and thus, a significant improvement of the AUC from the baseline (50,00) to the test set (66,20) can be seen (χ^2=0.02 , d.f.=1, p<0.001).

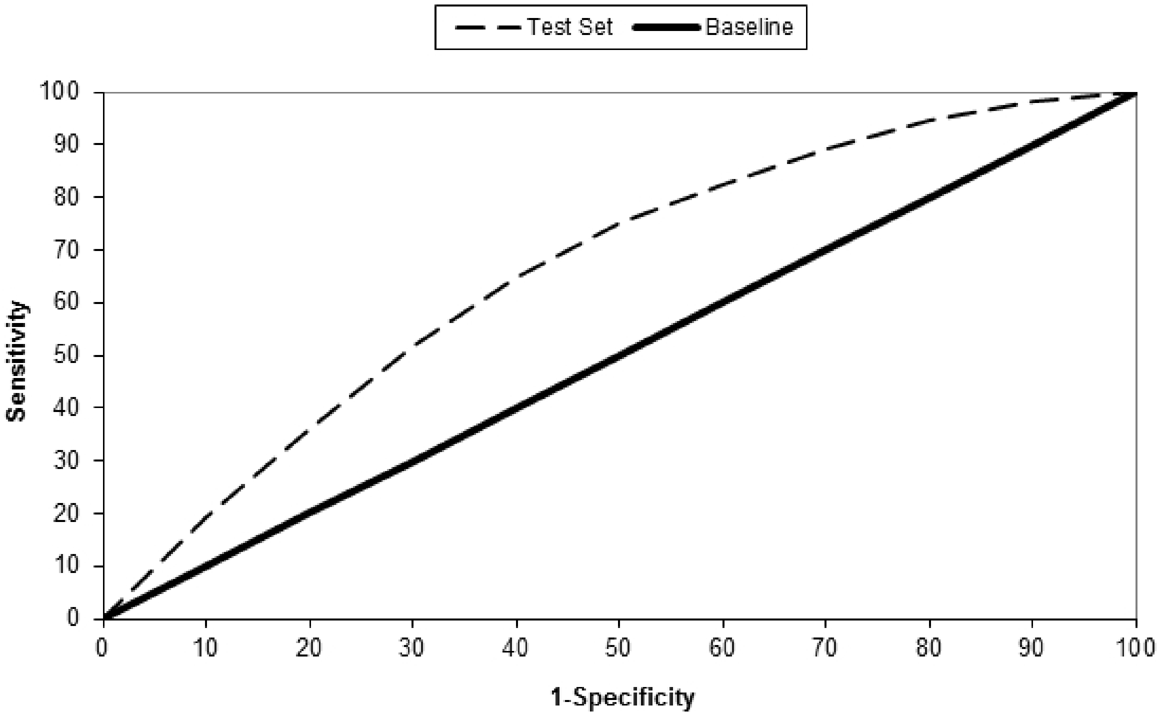

Figure 4: Sensitivity - specificity diagram of test set and baseline

4.4 Case study results

Examples for the technologies and the corresponding application fields extracted from the semantic textual patterns are presented below:

Meta-materials are materials with simultaneous negative dielectric constant and permeability. It seems that they contradict some optical laws because of their negative optical refractive index. The results of the case study show that projects applying meta-material technology in specific application fields are sensitive concerning espionage. Examples for applications are super-lenses that can be focused independent of light-wave length, antennas with improved emanation characteristics, suppression of parasitic waves in HF and microwave circuits, new types of optical and microwave elements (beam shapers, couplers, modulators, wave guides, resonators), and improved coatings for stealth aircraft.

In the field of semiconductor and LED laser technology, diode lasers are being developed in a wide spectral range (from ultraviolet to infrared) and with different emitting power. Examples for sensitive technologies and their applications are diode lasers operating in the mid-infrared range developed for optical countermeasures, short wavelength diode lasers for data storage, high-power diode lasers (emitting about 1 µm) for materials processing, and quantum cascade diode lasers with long wavelengths for sensor applications, including detection of hazardous substances and combat agents.

A further example is the infrared detector technology that is being developed for the atmospheric windows from 3 to 5 µm und 8 to 12 µm. New generations of infrared detectors are based on (Al-GaAs) quantum-well structures or on group-III antimonides. Sensitive projects apply this technology by creating a passive infrared detectors with high thermal and spatial resolutions with up to 640 x 512 pixels, by creating bi-spectral infrared detectors, and by creating thermal imaging cameras with two wavelengths for simultaneous detection (in both windows or in one) to permit an improved camouflage detection.

5 Conclusions

This work provides an automated approach based on text classification that calculates the information leakage risk of R&T projects that means the costs of an information loss and the probabilities that this loss will occur. Thus, it enables an automated identification of projects with high espionage risk. Literature introduces manual approaches where human experts evaluate the espionage risk of a few number of projects. However considering the large number of R&T projects in a country, in a specific application field, or in a research program this task only can be performed automatically for performance reasons.

Text classification is used because the identification of projects - that are a profitable target for espionage in contrast to projects that are not - depicts a binary classification model where a binary text classification approach can be applied. Based on these results, the risk can be calculated automatically for a further risk management process. This is in contrast to existing qualitative approaches.

The approach uses latent semantic indexing (LSI) to identify semantic textual patterns occurring within the textual information. These semantic textual patterns are used as variable in a binary prediction model to calculate the costs of a potential loss. Further, textual information about related technologies and application fields available in the internet are collected and integrated in this approach. This shows the competitive situation of a technology - application field combination and it enables the calculation of the probability that the loss will occur. Based on both calculations, the information leakage risk of R&T projects is estimated.

Overall, this approach is successful in the identification of projects with high espionage risk and it helps researchers, research planners, and governmental agencies to ensure the processing of risk reduction measures despite limited resources.

Acknowledgments
To realize this project, we use several software tools: SAS v9.1.3, SAS Text Miner v5.2, and Matlab v7.0.4. We also use a self-developed program for web mining.

References

Allison, P. D. (1999). *Logistic Regression using the SAS System: Theory and Application.* Cary: SAS Institute Inc.
Almeida, P. (1996). Knowledge sourcing by foreign multinationals: Patent citation analysis in the US semiconductor industry. *Strategic Management Journal*, 17, 155-165.
Brunnermeier, M. K. (2005). Information Leakage and Market Efficiency. *Review of Financial Studies*, 18(2), 417-457.
Buckinx, W., Moons, E., Van den Poel, D., & Wets, G. (2004). Customer-adapted coupon targeting using feature selection. *Expert Systems with Applications*, 26(4), 509-518.
Chen, M.-Y., Chu, H.-C., & Chen, Y.-M. (2010). Developing a semantic-enable information retrieval mechanism. *Expert Systems with Applications*, 37(1), 322-340.
Cherp, A., & Demidova, O. (2005). Risk assessment for improved treatment of health considerations in EI A. *Environmental Impact Assessment Review* ,25 (4), 411-429.
Choi, J. Y., Lee, J. H., & Sohn, S. J. (2009). Impact analysis for national R&D funding in science and technology using quantification method II. *Research Policy*, 38(10), 1534-1544.
Christidis, K., Mentzas, G., & Apostolou, D. (2012). Using latent topics to enhance search and recommendation in Enterprise Social Software. *Expert Systems with Applications*, 39(10), 9297- 9307.
Crane, A. (2005). In the company of spies: When competitive intelligence gathering becomes industrial espionage. *Business Horizons*, 48(3), 233-240.
Crawford, V., & Sobel, J. (1982). Strategic Information Transmission. *Econometrica*, 50(6), 1431-1451.
Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6), 391-407.
DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3), 837-845.

Finzen, J., Kintz, M., & Kaufmann, S. (2012). Aggregating web-based ideation platforms. *International Journal of Technology Intelligence and Planning*, 8(1), 32-46.
Fleck, J., & Howells, J. (2001). Technology, the Technology Complex and the Paradox of Technological Determinism. *Technology Analysis & Strategic Management*, 13(4), 523-531.
Gericke, W., Thorleuchter, D., Weck, G., Reiländer F., & Loß, D. (2009). Vertrauliche Verarbeitung staatlich eingestufter Information - die Informationstechnologie im Geheimschutz. *Informatik Spektrum,* 32(2), 102-109.
German Federal Ministry of the Interior (2011). Verfassungsschutzbericht der Bundesregierung (pp. 376-414). Berlin.
Geschka, H. (1983). Creativity techniques in product planning and development: A view from West Germany. *R&D Management,* 13(3), 169-183.
Geschka, H., Lenk, T., & Vietor, J. (2002). The idea and project database of WELLA AG. *International Journal of Technology Management,* 23(5), 410-416.
Gorman, M. E. (2002). Types of Knowledge and Their Roles in Technology Transfer. *The Journal of Technology Transfer*, 27(3), 219-231.
Greiff, W. R. (1998). A theory of term weighting based on exploratory data analysis. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, & J. Zobel (Eds.), *Proceedings of the 21st SIGIR.*
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29-36.
Herstatt, C., & Geschka, H. (2002). Need assessment in practice - methods, experiences and trends. *International Journal of Entrepreneurship and Innovation Management*, 2(1), 56-68.
Ho, S. J. (2007). An economic analysis of military intelligence. *Defence and Peace Economics*, 18(6), 485-493.
Ho, S. J. (2008). Extracting the Information: Espionage with Double Crossing. *Journal of Economcs*, 93(1), 31-58.
Jaffe, A. B. (1986). Technological Opportunity and Spillovers of R&D: Evidence from Firms' Patents, Profits and Market Value. *American Economic Review*, 76(5), 984-999.
Jiang, J., Berry, M. W., Donato, J. M., Ostrouchov, G., & Grady, N. W. (1999). Mining consumer product data via latent semantic indexing. *Intelligent Data Analysis*, 3(5), 377-398.
Jiménez, C. H. O., Garrido-Vega, P., Díez de los Ríos, J. L. P., & González, S.G. (2011). Manufacturing strategy-technology relationship among auto suppliers. *International Journal of Production Economics*, 133(2), 508-517.
Jones, A. (2008). Espionage. Industrial espionage in a hi-tech world. *Computer Fraud & Security*, 2008(1), 7-13.
Kaperonis, I. (1984). Industrial espionage. *Computers & Security*, 3(2), 117-121.
Kim, Y., Toh, K. A., Teoh, A .B. J., Eng, H. L., & Yau, W. Y. (2012). An online AUC formulation for binary classification. *Pattern Recognition*, 45(6), 2266-2279.
Ko, Y., & Seo, J. (2009). Text classification from unlabeled documents with bootstrapping and feature projection techniques. *Information Processing & Management*, 45(1), 70-83.
Lee, C. H., & Wang S. H. (2012). An information fusion approach to integrate image annotation and text mining methods for geographic knowledge discovery. *Expert Systems with Applications*, 39(10), 8954-8967.
Lee, S. C., Chang, S. N., Liu, C. Y., & Yang, J. (2007). The effect of knowledge protection, knowledge ambiguity, and relational capital on alliance performance. *Knowledge and Process Management*, 14(1), 58-69.
Lin, M-H., & Hong, C-F. (2011). Opportunities for Crossing the Chasm between Early Adopters and the Early Majority through New Uses of Innovative Products. *The Review of Socionetwork Strategies*, 5(2), 27-42.
Luo, Q., Chen, E., & Xiong, H. (2011). A semantic term weighting scheme for text categorization. *Expert Systems with Applications*, 38(10), 12708-12716.
Marhavilas, P. K., Koulouriotis, D., & Gemeni, V. (2011). Risk analysis and assessment methodologies in the work sites: On a review, classification and comparative study of the scientific literature of the period 2000-2009. *Journal of Loss Prevention in the Process Industries*, 24(5), 477-523.
Matsui, A. (1989). Information Leakage Forces Cooperation. *Games and Economic Behavior*, 1(1), 94-115.
Migueis, V. L., Van den Poel, D., Camanho, A. S., & Cunha, J.F. (2012). Modeling partial customer churn: On the value of first product-category purchase sequences. *Expert Systems with Applications*, 39(12), 11250-11256.
Oikonomou, I. (2012). The European Defence Agency and EU military space policy: Whose space odyssey?. *Space Policy*, 28(2), 102-109.
Orozco, D. (2012). Amending the Economic Espionage Act to Require the Disclosure of National Security-Related Technology Thefts. Catholic University Law Review, Forthcoming.
Park, D, H., Kim, H. K., Choi, I. Y., & Kim, J. K. (2012). A literature review and classification of recommender systems research. *Expert Systems with Applications*, 39(11), 10059-10072.
Prinzie, A., & Van den Poel, D. (2006). Investigating purchasing-sequence patterns for financial services using Markov, MTD and MTDg models. *European Journal of Operational Research*, 170(3), 710-734.

Prinzie, A., & Van den Poel, D. (2007). Predicting home-appliance acquisition sequences: Markov/Markov for Discrimination and survival analysis for modeling sequential information in NPTB models. *Decision Support Systems*, 44(1), 28-45.
Radder, H. (2009). Science, Technology and the Science-Technology Relationship. *Philosophy of Technology and Engineering Sciences*, 2009, 65-91.
Reisman, A. (2006). A taxonomic view of illegal transfer of technologies: A case study. *Journal of Engineering and Technology Management*, 23(4), 292-312.
Rubenstein, A. H., Douds, C. F., Geschka, H., Kawase, T., Miller, J. P., Saintpaul, R., & Watkins, D. (1977). Management perceptions of government incentives to technological innovation in England, France, West Germany and Japan. *Research Policy*, 6(4), 324-357.
Salton, G., Allan, J., & Buckley, C. (1994). Automatic structuring and retrieval of large text files. *Communications of the ACM*, 37(2), 97-108.
Shi, L., & Setchi, R. (2012). User-oriented ontology-based clustering of stored memories. *Expert Systems with Applications*, 39(10), 9730-9742.
Si, H., Ji, H., & Zeng, X. (2012). Quantitative risk assessment model of hazardous chemicals leakage and application. *Safety Science*, 50(7), 1452-1461.
Sivanesan, G. (2011). The human factor in espionage. *Computer Fraud & Security*, 2011(2), 15-16.
Solan, E., & Yariv, L. (2004). Games with espionage. *Games and Economic Behavior*, 47(1), 172-199.
Subramanian, A. M., & Soh, P. H. (2010). An empirical examination of the science-technology relationship in the biotechnology industry. *Journal of Engineering and Technology Management*, 27(3-4), 160-171.
Sudhamathy, G. & Jothi Venkateswaran, C. (2012). Fuzzy Temporal Clustering Approach for E-Commerce Websites. *International Journal of Engineering and Technology*, 4(3), 119-132.
Te Kulve, H., & Smit, W. A. (2003). Civilian-military co-operation strategies in developing new technologies. *Research Policy*, 32(6), 955-970.
Thorleuchter, D. (2008). Finding Technological Ideas and Inventions with Text Mining and Technique Philosophy. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data Analysis, Machine Learning, and Applications* (pp. 413-420). Berlin: Springer.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010a). Mining Ideas from Textual Information. *Expert Systems with Applications*, 37(10), 7182-7188.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010b). A compared R&D-based and patent-based cross impact analysis for identifying relationships between technologies. *Technological Forecasting and Social Change*, 77(7), 1037-1050.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010c). Mining innovative ideas to support new product research and development. In H. Locarek-Junge, & C. Weihs (Eds.), *Classification as a Tool for Research* (pp. 587-594). Berlin: Springer.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2010d). Extracting consumers needs for new products - A web mining approach. In *Proceedings WKDD 2010* (p. 441.). Los Alamitos: IEEE Computer Society.
Thorleuchter, D., & Van den Poel, D. (2011a). Semantic Technology Classification - A Defence and Security Case Study. In Proc. *Uncertainty Reasoning and Knowledge Engineering* (pp. 36-39). New York: IEEE.
Thorleuchter, D., & Van den Poel, D. (2011b). Companies Website Optimising concerning Consumer's searching for new Products. In Proc. Uncertainty Reasoning and Knowledge Engineering (pp. 40-43). New York: IEEE.
Thorleuchter, D., & Van den Poel, D. (2011c). High Granular Multi-Level-Security Model for Improved Usability. In: System Science, Engineering Design and Manufacturing Informatization 1 (pp. 191-194). New York: IEEE.
Thorleuchter, D., Herberz, S., & Van den Poel, D. (2012). Mining Social Behavior Ideas of Przewalski Horses. *Lecture Notes in Electrical Engineering*, 121, 649-656.
Thorleuchter, D., Schulze, J., & Van den Poel, D. (2012). Improved Emergency Management by Loosely Coupled Logistic System. *Communications in Computer and Information Science*, 318, 5-8.
Thorleuchter, D., Van den Poel, D., & Prinzie, A. (2012). Analyzing existing customers' websites to improve the customer acquisition process as well as the profitability prediction in B-to-B marketing. *Expert Systems with Applications*, 39(3), 2597-2605.
Thorleuchter, D., & Van den Poel, D. (2012a). Extraction of Ideas from Microsystems Technology. *Advances in Intelligent and Soft Computing*, 168, 563-568.
Thorleuchter, D., & Van den Poel, D. (2012b). Predicting e-commerce company success by mining the text of its publicly-accessible website. *Expert Systems with Applications*, 39(17), 13026-13034.
Thorleuchter, D., & Van den Poel, D. (2012c). Using NMF for Analyzing War Logs. *Communications in Computer and Information Science*, 318, 73-76.

Thorleuchter, D., & Van den Poel, D. (2012d). Using Webcrawling of Publicly-Available Websites to Assess E-Commerce Relationships. In *SRII Global Conference* 2012 (pp. 402-410). San Jose, CA, USA: IEEE.

Thorleuchter, D., & Van den Poel, D. (2012e). Improved Multilevel Security with Latent Semantic Indexing. *Expert Systems with Applications*, 39(18), 13462-13471.

Thorleuchter, D. & Van den Poel, D. (2012f). Espionage Risk Assessment for Security of Defense based Research and Technology. *In: The 36th Annual Conference of the German Classification Society (GfKl) on Data Analysis, Machine Learning and Applications*, (p.128). Hildesheim: University of Hildesheim.

Thorleuchter, D., Weck, G., & Van den Poel, D. (2012a). Granular Deleting in Multi Level Security Models - an Electronic Engineering approach. Lecture Notes in Electrical Engineering, 1, 177, 609-614.

Thorleuchter, D., Weck, G., & Van den Poel, D. (2012b). Usability based Modeling for Advanced IT-Security - an Electronic Engineering approach. Lecture Notes in Electrical Engineering, 1, 177, 615-619

Thorleuchter, D., & Van den Poel, D. (2013). Technology classification with latent semantic indexing. *Expert Systems with Applications*, in Press.

Thorleuchter, D. (2004). Vorhabenbewertung im Rahmen der Rüstungsabschirmung. Euskirchen: INT-Report 193.

Tsai, H. H. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. *Expert Systems with Applications*, 39(9), 8172-8181.

Van den Poel, D., De Schamphelaere, J., & Wets, G (2004). Direct and indirect effects of retail promotions on sales and profits in the do-it-yourself market. *Expert Systems with Applications*, 27(1), 53-62.

Van Erkel, A. R., & Pattynama, P. M. T. (1998). Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. *European Journal of Radiology*, 27(2), 88-94.

Warner, W. T. (1994). International technology transfer and economic espionage. *International Journal of Intelligence and Counter Intelligence*, 7(2), 143-160.

Whitney, M. E., & Gaisford, J. D. (1999). An Inquiry Into the Rationale for Economic Espionage. *International Economic Journal*, 13(2), 103-123.

Yu, L., Hurley, T., Kliebenstein, J., & Orazem, P. (2012). A test for complementarities among multiple technologies that avoids the curse of dimensionality. *Economics Letters*, 116(3), 354-357.

Yucel, G., Cebi, S., Hoege, B., & Ozok, A. F. (2012). A fuzzy risk assessment model for hospital information system implementation. *Expert Systems with Applications*, 39(1), 1211-1218.

Zeng, J., Duan, J., Cao, W., & Wu, C. (2012). Topics modeling based on selective Zipf distribution. *Expert Systems with Applications*, 39(7), 6541-6546.

Zhong, J., & Li, X. (2010). Unified collaborative filtering model based on combination of latent features. *Expert Systems with Applications*, 37(8), 5666-5672.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort.* Cambridge: Addison-Wesley.